ESD-TR-67-237

Technical Note 1967-9

Vocoded Speech in the Absence of the Laryngeal Frequency

A. J. Goldberg
B. Gold, Editor

3 April 1967

Prepared under Electronic Systems Division Contract AF 19(628)-5167 by

Lincoln Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Lexington, Massachusetts

The work reported in this document was performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology, with the support of the U.S. Air Force under Contract AF 19(628)-5167. Certain portions of the work were supported under other contracts.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LINCOLN LABORATORY

VOCODED SPEECH IN THE ABSENCE OF THE LARYNGEAL FREQUENCY

A. J. GOLDBERG

Massachusetts Institute of Technology

B. GOLD, Editor

Group 62

TECHNICAL NOTE 1967-9

3 APRIL 1967

LEXINGTON MASSACHUSETTS

ABSTRACT

Most pitch excited channel vocoders require the fundamental or laryngeal frequency of the input speech to be present if the output speech is to be of high quality. In order to determine if speech whose fundamental is absent can have its pitch accurately restored so as to be used as an input to a vocoder, a computer simulation was performed. The fundamental was restored by passing the speech through a fullwave rectifier followed by a slope filter. The accuracy of the pitch restoration of this method was compared with that of simply measuring the pitch of speech whose fundamental was present by slope filtering alone. A third pitch detection method, that of visually displaying the speech waveform and determining the pitch by eye, was also used as a comparison. Pitch contours of the three methods indicate that pitch restored by fullwave rectification and slope filtering has larger perturbations than pitch as detected by slope filtering alone. Both methods produced pitch contours having much larger perturbations than pitch determined visually.

Speech whose pitch was determined by the above three methods was used to excite the spectrally flattened Lincoln Laboratory Vocoder. Listening tests of the vocoder output indicate that pitch restored by fullwave rectification and slope filtering produced rougher sounding speech than speech whose pitch was detected by slope filtering alone, but both methods produced speech having considerably more audible roughness than that produced by visually detected pitch. Finally, the sophisticated pitch detector of the vocoder itself produced speech of quality comparable to that determined visually.

Accepted for the Air Force
Franklin C. Hudson
Chief, Lincoln Laboratory Office

CONTENTS

VOCODED SPEECH IN THE ABSENCE OF THE LARYNGEAL FREQUENCY

I. INTRODUCTION

Recent interest in speech bandwidth compression devices has renewed the search for accurate methods of extraction of the laryngeal frequency or pitch from human speech. One of the most widely known speech bandwidth compression devices is the vocoder,[1,2,3,4] invented by Homer Dudley in the 1930's. To effect a bandwidth compression of the speech, the vocoder uses the fact that speech, at least in the steady state, has a frequency spectrum consisting of a fundamental and harmonics of this fundamental (see Fig. 1). The successful operation of the vocoder depends critically on the accurate detection of this fundamental, or pitch, of the input speech. Early vocoder attempts to compress the bandwidth of speech produced speech of unreliable quality but recent attempts have shown that this reliability can be improved. To a large degree, this improvement in vocoded speech can be traced to the development of devices that more accurately detect pitch[5,6,7,8,9] than those previously employed.

Many pitch detection schemes require that the speech fundamental be present in the input spectrum to the vocoder. However, sometimes this fundamental is filtered out so that the input speech to the vocoder has no fundamental present. For example, the speech fundamental generally is between 50-400 Hz,[10] but telephone lines have a lower cutoff of about 300 Hz. Hence,the speech at the telephone central office, which is where the vocoder could most profitably be situated, may often not contain a fundamental.

II. THE PURPOSE OF THE THESIS

The primary intent of this paper is to examine the possibility of reconstruction of an accurate speech fundamental where the original has been removed as in the above example. If this can be done, it will then be possible to detect the pitch in speech with the fundamental absent by first restoring the fundamental and then using present pitch detection techniques.

McKinney[11] has hypothesized that steady state speech can have its fundamental restored if it has previously been removed if

$$\sum_{n=2}^{M} nA_n \leq A_1 \qquad (1)$$

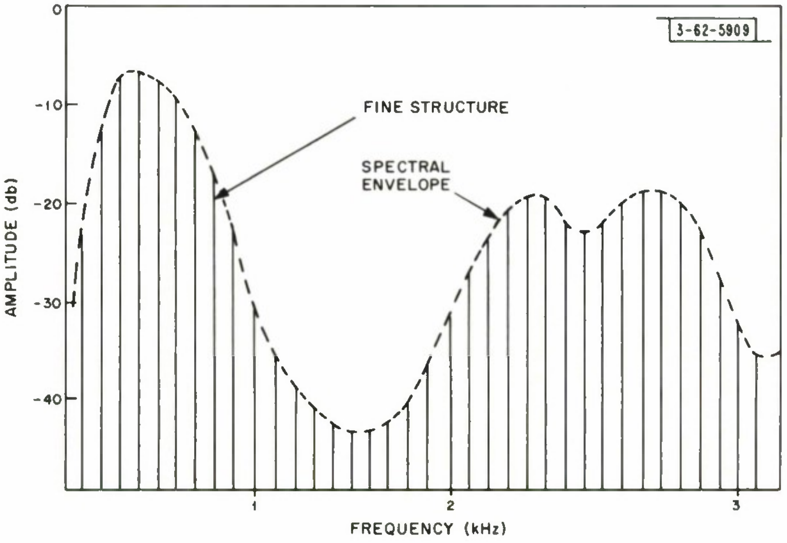

Fig. 1. Logarithmic Spectrum of "ee" Sound in S<u>ee</u>.

where n = the harmonic number

A_n = the amplitude of the n^{th} harmonic

M = the highest harmonic present in the filtered speech

If this result is not satisfied then McKinney states that accurate construction of the fundamental cannot be assured. He states that a computer simulation has validated this result for some special cases and that actual measurement of the fundamental can be made by passing the speech through a non-linear device. The pitch can then be obtained by applying conventional pitch extraction techniques to the output. Mathematical analysis seems to indicate that a full wave rectifier would be best although any even power non-linear device should perform well, assuming, of course, that the filtered speech satisfies Eq. (1).

Speech is not completely of a steady state nature and McKinney's analysis does not cover the situations where the pitch is varying in time. To determine what happens when speech has its fundamental removed and then restored by a non-linear device a computer simulation was performed and the results described in this paper. These results are compared with the pitch detected by several methods wherein the speech fundamental was present. The pitch data of the various pitch detection schemes along with the original speech was used to excite a vocoder and subjective listening tests performed on the outputs to determine what effect reconstruction of the fundamental has on speech quality.

III. THE EXPERIMENT

The first part of the experiment is a computer simulation of three pitch detection schemes to determine the effects on the pitch of speech whose fundamental has been filtered out but has later been restored. The computer used in this simulation was MIT's Electrical Engineering Department's PDP-1. For a brief description of some of the features of this facility see Appendix 1.

A. Speech Input to Computer

In order to detect pitch by computer it was first necessary to convert the analog speech signal into numerical data so that it could be stored and manipulated by the computer. This was performed using an 8-bit analog to digital (A-D) converter with timing provided by the computer program and the PDP-1. For a more complete description of this operation see Appendix 2. The A-D converter sampled the speech at a 10 kHz rate and computer memory space was sufficient to store 1-1/2 seconds of sampled speech at this rate.

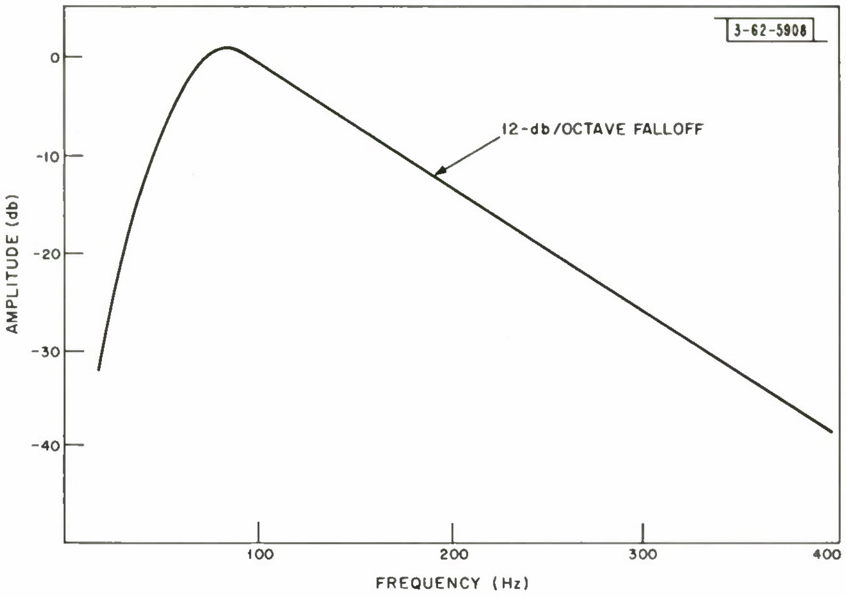

Fig. 2. Frequency Response of the Ideal Slope Filter With 12 db/Octave Falloff Above 80 Hz.

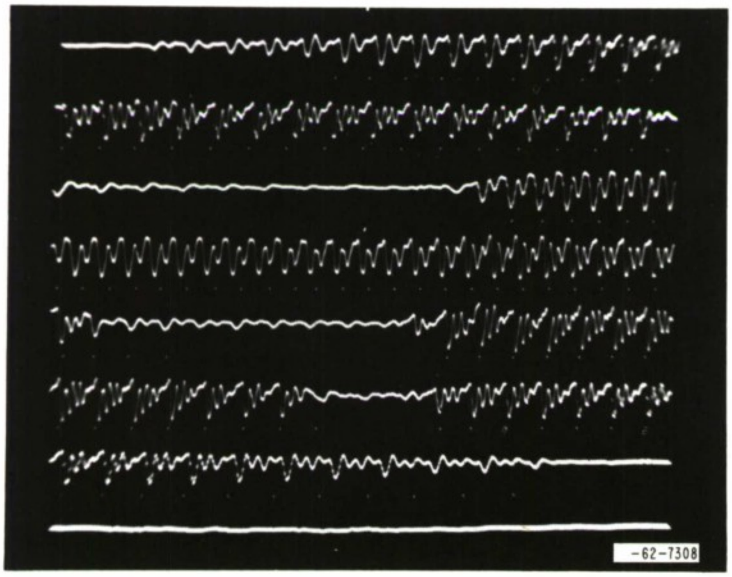

Fig. 3. 0-3000 Hz Speech Signal (150 ms/Line) of "We Are Due About Eight" – Speaker 1.

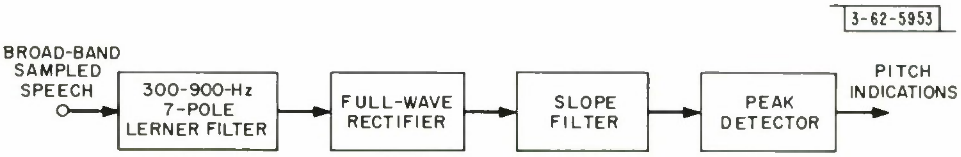

Fig. 4. Pitch Detection in the Absence of the Fundamental Frequency.

B. Description of Three Pitch Detection Techniques

1. Pitch Detection in the Absence of the Fundamental

To detect pitch when the fundamental was not present an approach illustrated in Fig. 4 was used. First, to assure that no fundamental was present, the speech data was filtered by a 300-900 Hz 7-pole Lerner filter.[12] For a discussion of the digital filters used in this experiment see Appendix 3. The resulting 300-900 Hz waveform was then fullwave rectified to restore the fundamental and then sent through a slope filter having a slope of 12 db/octave beginning at about 80 Hz as illustrated in Fig. 2. Since the fundamental generally was between 50 and 400 Hz, this slope filter would enhance the fundamental with respect to its harmonics, provided the pitch was above 80 Hz. As a result, the number of local maxima per second of the output of the slope filter was usually proportional to the pitch of the speech wave. *

The output of this slope filter was then examined for local maxima. The time differences between these maxima represented the local period of the speech wave. The original speech wave and the pitch periods were then displayed on the scope face of the computer in a manner similar to that illustrated in Fig. 3. Any data points that represented gross errors like frequency doubling were eliminated. This speech data was then recorded on audio tape in a method described in section III C.

2. Pitch Detection With a Slope Filter in the Presence of a Fundamental

As a comparison against the detection scheme described above, a second pitch detection technique was used in which the speech fundamental was present (see Fig. 5). In this scheme the 300-900 Hz filter and fullwave rectifier were omitted. The speech signal was sent through the same slope filter used above, followed by the local maximum detector. Again the original speech waveforms plus the pitch periods were visibly displayed and gross pitch errors were removed. The speech and pitch indications were then recorded on tape as described in section III C.

*However, this was not always the case. For example, if the speech signal had a first harmonic that was 12 db above the fundamental then the output of the slope filter would have the fundamental and first harmonic of equal energies and the number of local maximums/second of this output would then not be proportional to the fundamental alone.

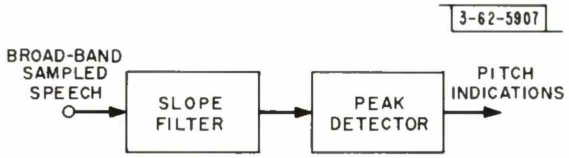

Fig. 5. Pitch Detection by Slope Filtering With the Fundamental Present.

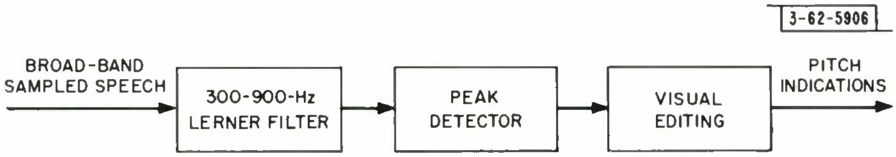

Fig. 6. Pitch Detection by Visual Means.

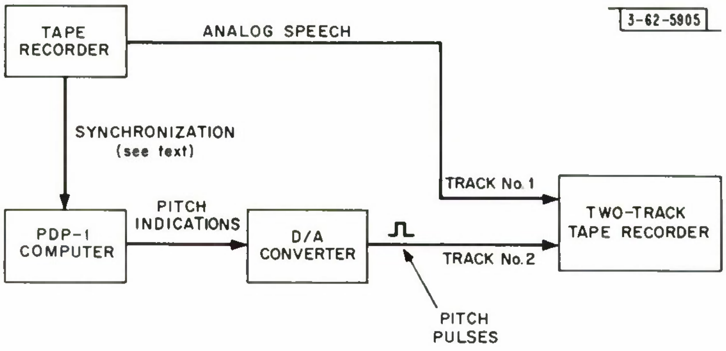

Fig. 7. System for Recording Speech and Pitch Pulses on a Two Track Audio Tape.

3. Pitch Detection by Eye with the Fundamental Present

As a further check against both of the above methods the speech was visibly displayed[6] and the periods recorded in the computer memory. To facilitate this, the original speech was first filtered by using the 300-900 Hz Lerner filter described earlier (see Fig. 6). Next, distances between local maximum were recorded with the aid of the local maximum locator used in the other two pitch detection schemes. These maximum points and the filtered wave were displayed and all indications of local maximum that did not correspond to periods of this wave were discarded. These apparently correct pitch indications and the speech were then recorded on audio tape in the method described below.

C. Read Out of Pitch Data and the Lincoln Laboratory Vocoder

The final result of the computer simulations described above was a two-track audio tape recording. One track contained the original analog speech while the second track contained analog pitch pulses that were synchronized with the analog speech (see Fig. 7 and Appendix 2). This two-track recording was then used as input to the Lincoln Laboratory vocoder,[4] which contains spectral flattening in its synthesizer stage. The analog speech was fed into the analyzer stage of the vocoder in the conventional manner while the analog pitch pulses were fed into the synthesizer stage of the vocoder as shown in Fig. 8. Thus, by disconnecting the pitch detector of the vocoder, the pitch pulses generated by the computer simulation could be substituted for those ordinarily generated by the vocoder itself. In practice the filters of the analyzer stage delayed the speech with respect to the pitch by approximately 10 ms. It was felt that this was of only minor importance in the experiment and consequently no attempt was made to correct this by also delaying the pitch pulses. Audio tape recordings were made of the vocoder outputs for each of the three pitch detection schemes described above. For reference, recordings were also made of the vocoder output with excitation derived from its own pitch extractor.

IV. RESULTS

Three sentences, each of about one second's duration, were used to compare the various pitch detection schemes. The sentences used were:

1. We are due about eight.
2. Whom am I to meet?
3. I love you!

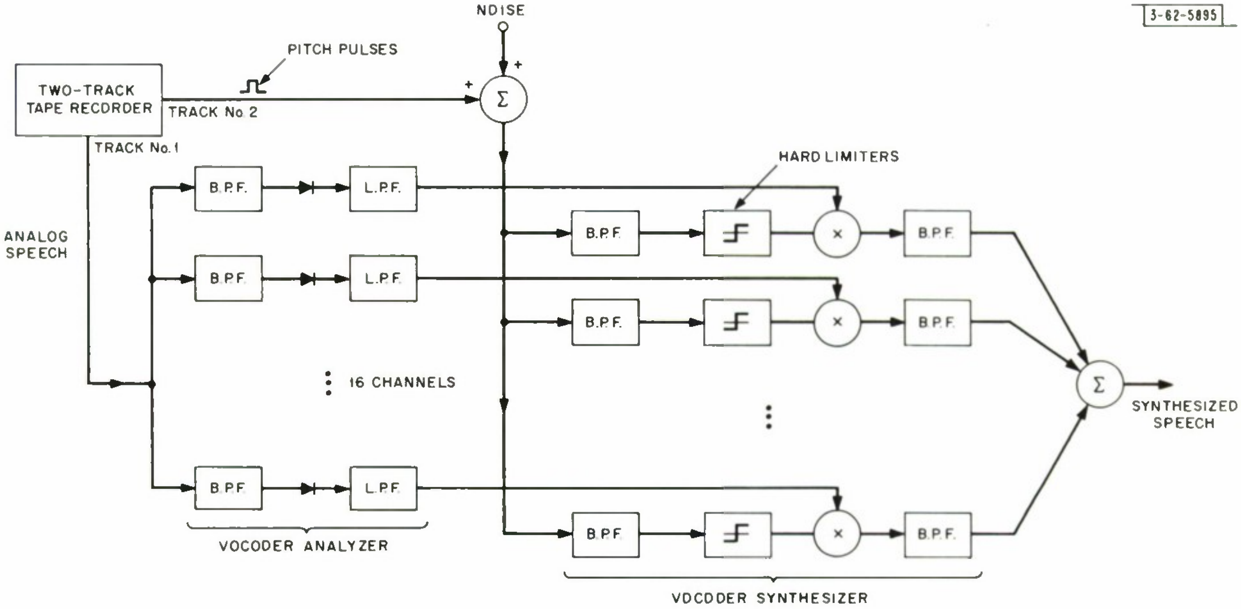

Fig. 8. Construction of Synthesized Speech Employing Computer Detected Pitch and the Lincoln Laboratory Vocoder.

Each sentence was spoken once by three different speakers and recorded on audio tape. It was this tape that was used as input to the computer and to one track of the two-track tape recording used to excite the Lincoln Laboratory vocoder.

A new audio tape recording was made of the vocoder output and subjective tests were performed. This audio tape recording was made up of groups of two closely spaced sentences. Each of the two sentences were identical (i.e. the same speaker and same words) except for the pitch detection scheme used to excite the vocoder. The pitch detection schemes were

A. Construction of the fundamental by fullwave rectification and slope filtering (fundamental originally absent).

B. Slope filtering with the fundamental present.

C. Eye detection of pitch.

D. Normal pitch detection by the vocoder's pitch detector.

Ten (10) subjects were chosen to listen to the recorded vocoder outputs. These ten listeners included people who have very little knowledge of speech mechanisms and those who can discern subtle differences in speech quality. The listeners were asked which sentence in each group of the two sentences they thought sounded less rough in quality or, if both were equally rough, which they preferred. They could also indicate that they had no preference between the two sentences. This comparison was performed on sixty-eight (68) two sentence combinations.

The following comparisons of pitch detection schemes were made for each of three sentences and each of three speakers.

Comparison of Pitch Detection Techniques

Comparison 1

(A) fundamental absent, fullwave rectification and slope filtering compared to

(B) slope filter-fundamental present

Comparison 2

(A) fundamental absent-fullwave rectification and slope filtering compared to

(D) vocoder detected pitch-fundamental present

Comparison 3

(B) slope filter-fundamental present compared to

(D) vocoder detected pitch-fundamental present

Comparison 4

(C) eye detected pitch compared to

(D) vocoder detected pitch-fundamental present

For each comparison of the form type X compared with type Y a second comparison at a later point in the test of the form type Y compared with type X was done. This tended to cancel effects of one pitch detection scheme appearing first in a comparison and the other second as the reverse comparison was also made.

The results appear below for ten subjects. One of the two-track tape recordings of sentence 3, speaker 1 comparison B was accidentally destroyed so this comparison could not be made.

Comparison of Four Pitch Detection Techniques

# times preferred	# times preferred	# times preferred
A to B	B to A	neither
44	69	47
# times preferred	# times preferred	# times preferred
A to D	D to A	neither
28	108	44
# times preferred	# times preferred	# times preferred
B to D	D to B	neither
20	101	39
# times preferred	# times preferred	# times preferred
C to D	D to C	neither
42	81	57

(See text for meanings of comparisons A, B, C, D).

$$\frac{\text{\# times preferred B}}{\text{\# times preferred A}} = 1.57$$

$$\frac{\text{\# times preferred D}}{\text{\# times preferred A}} = 3.84$$

$$\frac{\text{\# times preferred D}}{\text{\# times preferred B}} = 5.05$$

$$\frac{\text{\# times preferred D}}{\text{\# times preferred C}} = 1.95$$

All tests involving the comparison of the pitch detector of the vocoder with another technique showed that the sophisticated pitch detector of the vocoder produced speech of quality superior to any other pitch detection scheme used in the experiment. It also showed that eye detected pitch was better than either of the other two computer generated pitch schemes by a factor of almost 2 to 1.

The comparison also showed that the listeners preferred vocoded speech that resulted from speech whose pitch was present and detected by slope filtering techniques to speech whose pitch was restored by full-wave rectification and then slope filtered and detected. Thus, it appears that the pitch resulting from artificial construction of the fundamental does not produce speech of quality comparable to that produced by pitch detected from speech whose fundamental has not been previously destroyed. Hence, an attempt to reconstruct an accurate fundamental by full wave rectification does not appear to have been successful.

The listening comparisons also indicated that the slope filter technique caused a considerable amount of roughness in speech quality. This was apparent not only in the listening comparisons made of the ten subjects but on a great many other speech samples as well. It is felt that by prefiltering the speech the peaks corresponding to the periods of the resulting waveform are too broad to accurately determine pitch. The actual computer generated pitch contours for each of the three sentences spoken by the three speakers used in the experiment were plotted in Fig. 9 through Fig. 17. A comparison of pitch data detected by a slope filter to that detected by eye showed that the former method tended to generate pitch contours having large pitch perturbations while the latter method produced contours having smaller pitch perturbations. Since the only actual difference between the two methods was a prefiltering of the speech by a slope filter it must be concluded that the slope filter caused the roughness in pitch.

The same graphs illustrated that reconstruction of the fundamental by full wave rectification followed by slope filtering also produced large perturbations in pitch. A comparison of the pitch contours indicated that those contours resulting from full-wave rectification and slope filtering had larger perturbations that those resulting from slope filtering alone. Thus, it seemed that artificial regeneration of the speech fundamental had resulted in a fundamental that had some unwanted jitter associated with it. This result seems to support the results of R. R. Reisz.[13]

Reisz states that the first harmonic is not always the vibrating frequency of the vocal cords. If the frequency of vibration of the vocal cords is maintained constant and the physical shape of the resonant cavities of the vocal tract is also maintained constant (as in a sustained vowel sound) then the frequency of the n^{th} harmonic at the the mouth opening is the same as that of the vocal cord wave. However, if the physical shape of the resonant cavities of the vocal tract is changing in time such as during a sequence of speech sounds, then the resulting changes in the phase shift along the vocal tract give rise to an apparent frequency of the harmonic component which is somewhat different from its frequency in the steady state. Similarly, harmonic frequencies are not integral multiples during pitch variation. Therefore, any attempt to construct the fundamental by a detection arrangement that uses just the differences in frequency of harmonic components will result in a fundamental that differs from the true fundamental. This will be true because the harmonics are not strictly multiples of the fundamental since the resonant cavities of the vocal tract are physically changing. Consequently, correct fundamentals can be determined from harmonics in steady state speech but not during transition periods. This reasoning would also account for the fact that the pitch contours of the full wave rectification detection scheme vary more rapidly than that of pitch contours generated by slope filtering alone. Therefore, McKinney's analysis that predicts the conditions for which the fundamental can be detected, although correct for steady state speech, does not apply during a sequence of speech sounds. His analysis does not include the fact that in moments of transition the harmonics are not exact multiples of the fundamental, but are also a function of the physically changing vocal tract, and vocal cords.

V. CONCLUSION

A computer simulation and vocoder tests employing a vocoder with special flattening have shown that if speech with a missing fundamental has the fundamental restored by full-wave rectification then this speech will produce vocoded speech having an audible roughness. Surprisingly, if the original speech to the vocoder contains the fundamental and pitch detection is performed by slope filtering techniques, then the vocoder outputs produce speech containing a considerable amount of audible roughness. Pitch as determined by eye, however, produced more natural sounding speech than either of the above methods of pitch extraction. Pitch contours indicate that pitch as determined by eye has fewer pitch perturbations than pitch that has either been restored by full-wave rectification or has been detected by slope filtering. These contours indicate that there is a correlation between the audible roughness of vocoded speech and the visible roughness of the pitch contours.

Finally the sophisticated pitch extractor of the Lincoln Laboratory vocoder operating on speech containing a fundamental produced a naturalness in speech comparable to that produced by eye detected pitch. Both the Lincoln Laboratory pitch extractor and pitch extraction by eye produced speech containing much less audible roughness than either slope filtering with the fundamental present or artificial reconstruction of the fundamental by full-wave rectification.

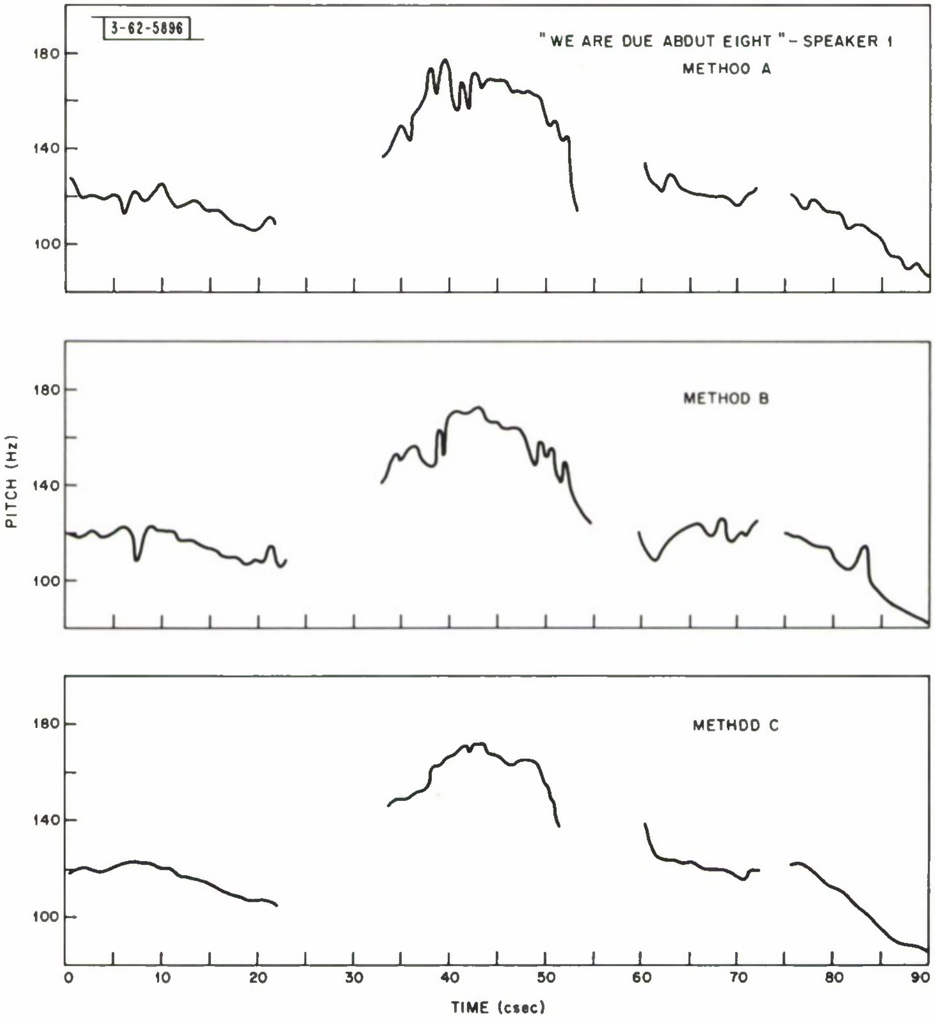

Fig. 9. Pitch Contours of Sentence "We Are Due About Eight" – Speaker 1.

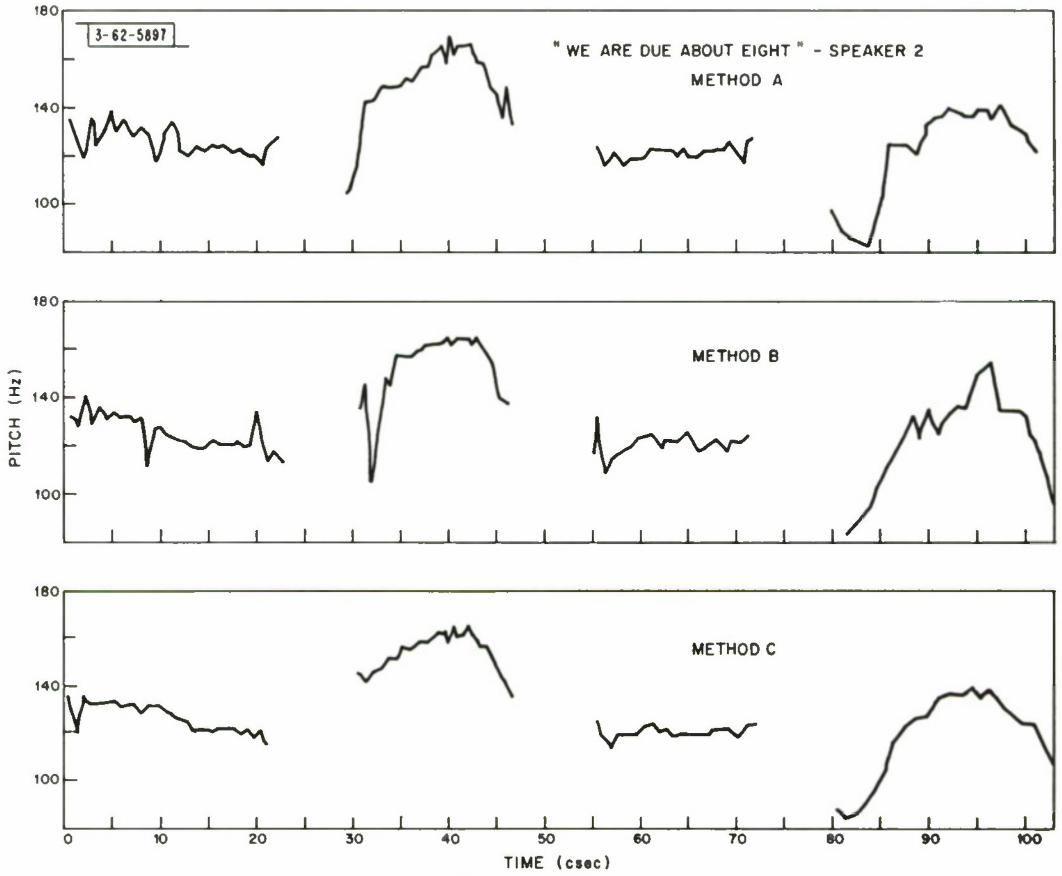

Fig. 10. Pitch Contours of Sentence "We Are Due About Eight" – Speaker 2.

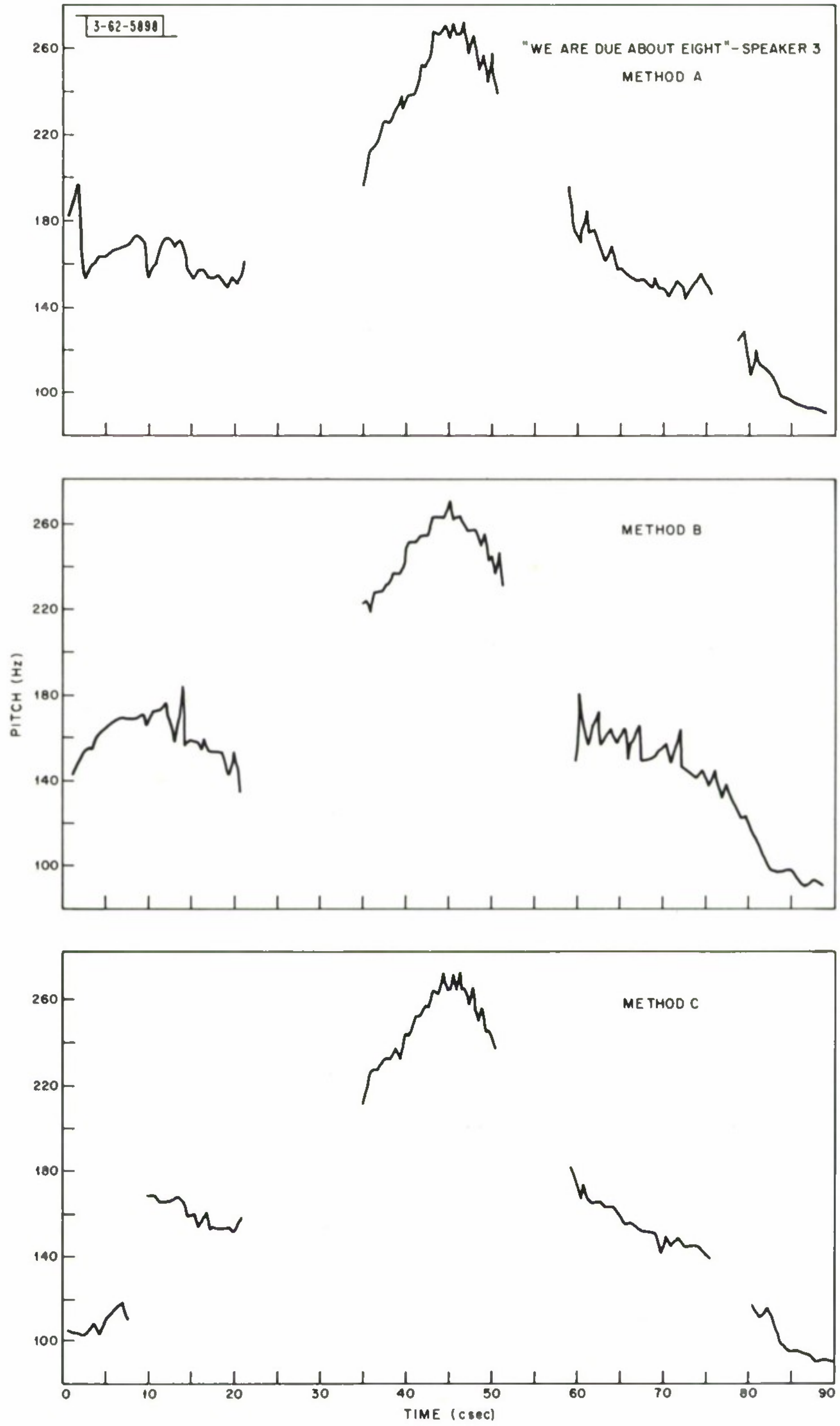

Fig. 11. Pitch Contours of Sentence "We Are Due About Eight" – Speaker 3.

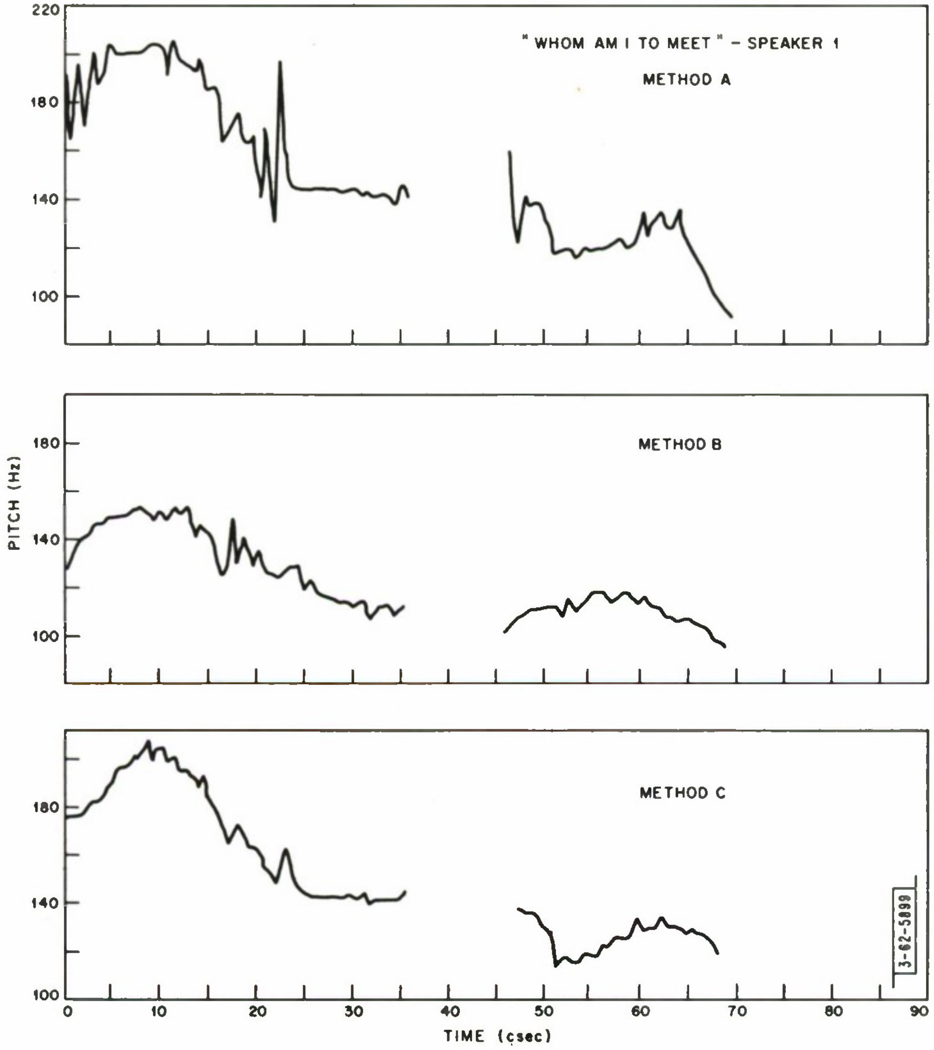

Fig. 12. Pitch Contours of Sentence "Whom Am I To Meet" – Speaker 1.

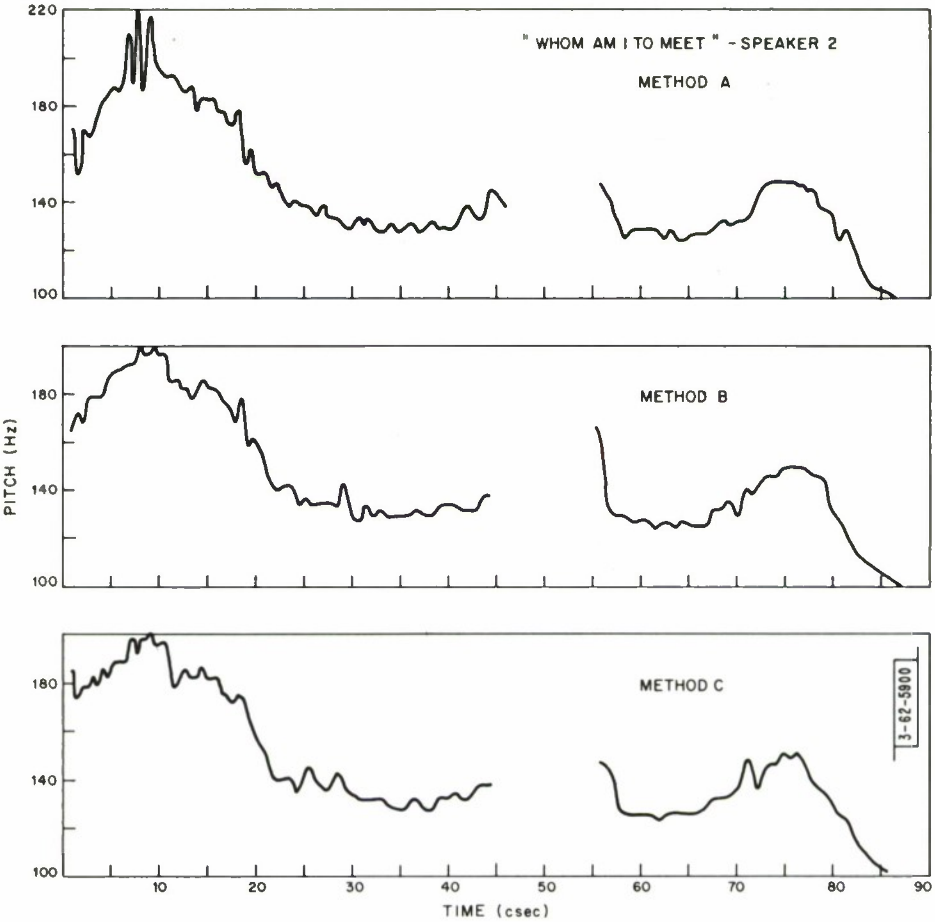

Fig. 13. Pitch Contours of Sentence "Whom Am I To Meet" – Speaker 2.

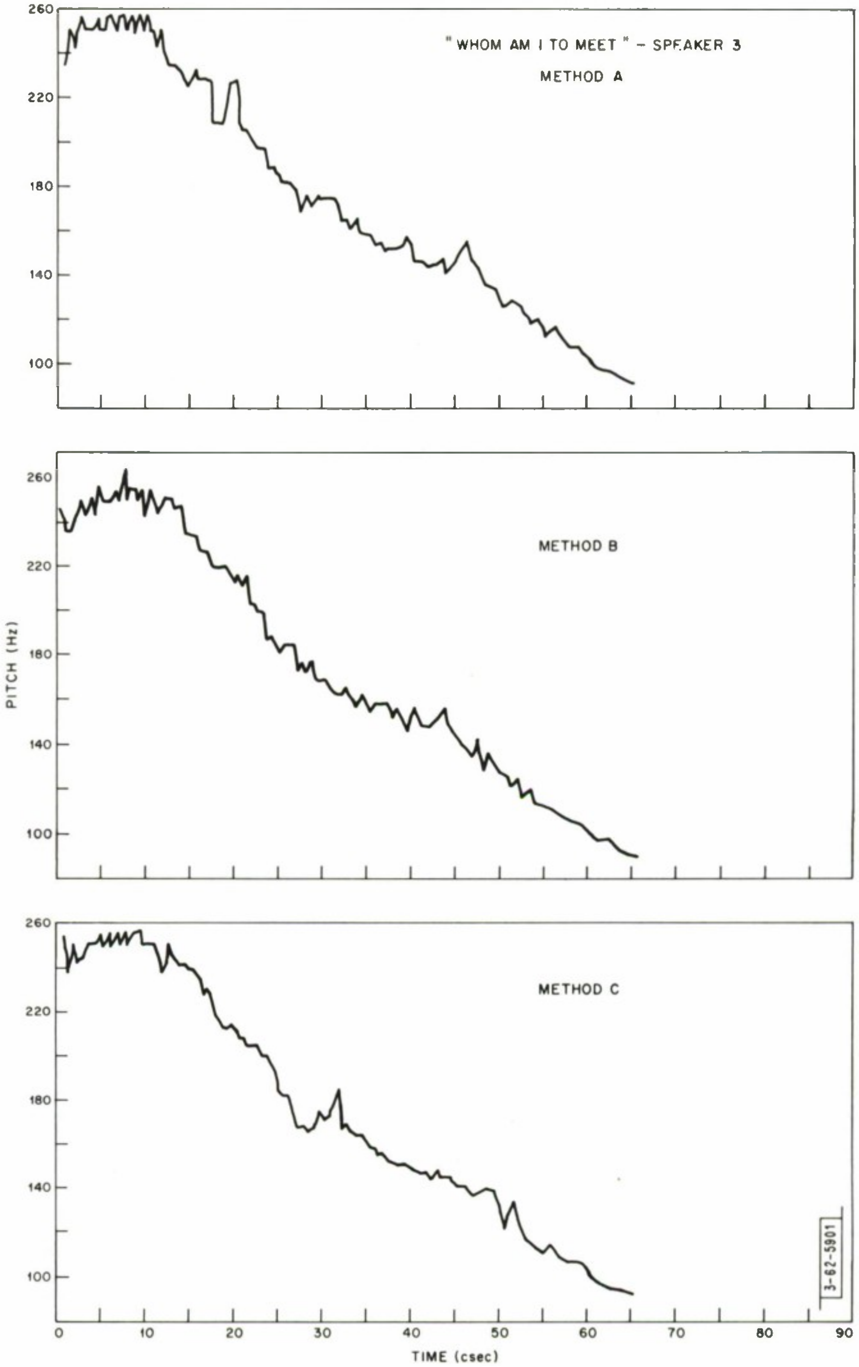

Fig. 14. Pitch Contours of Sentence "Whom Am I To Meet" – Speaker 3.

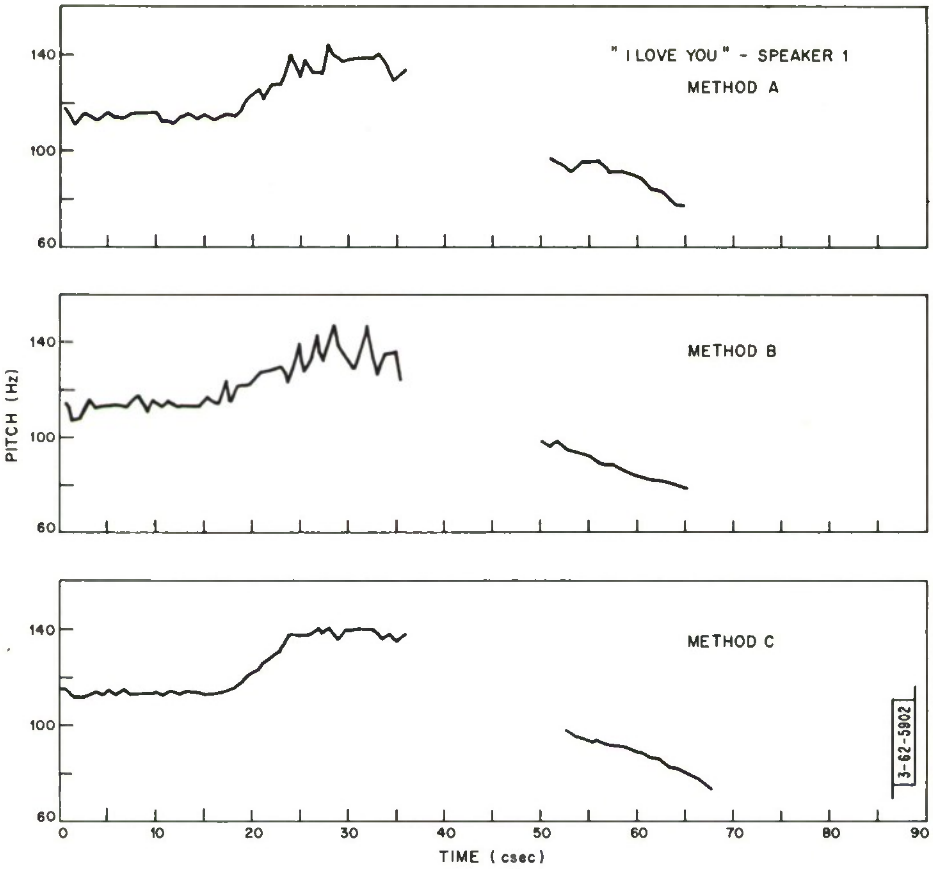

Fig. 15. Pitch Contours of Sentence "I Love You." – Speaker 1.

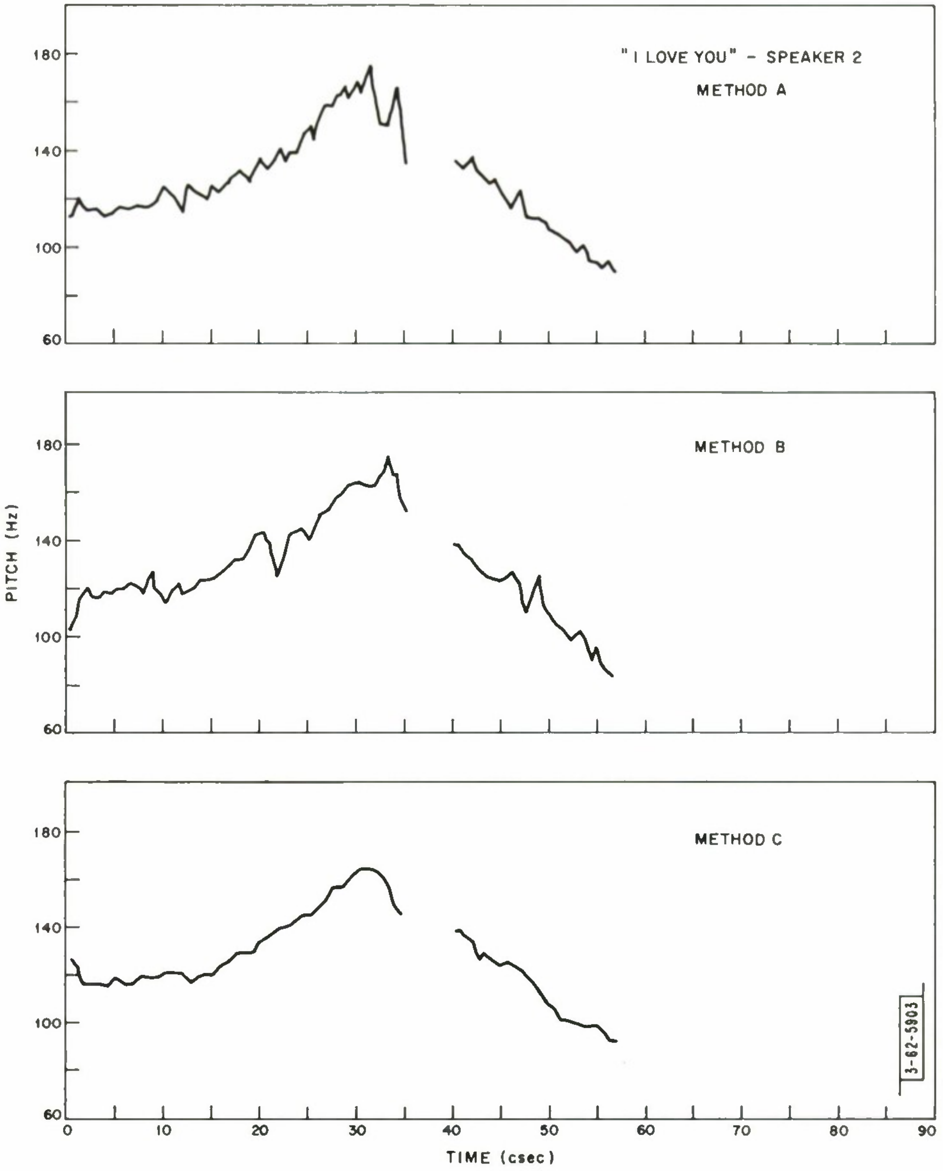

Fig. 16. Pitch Contours of Sentence "I Love You!" – Speaker 2.

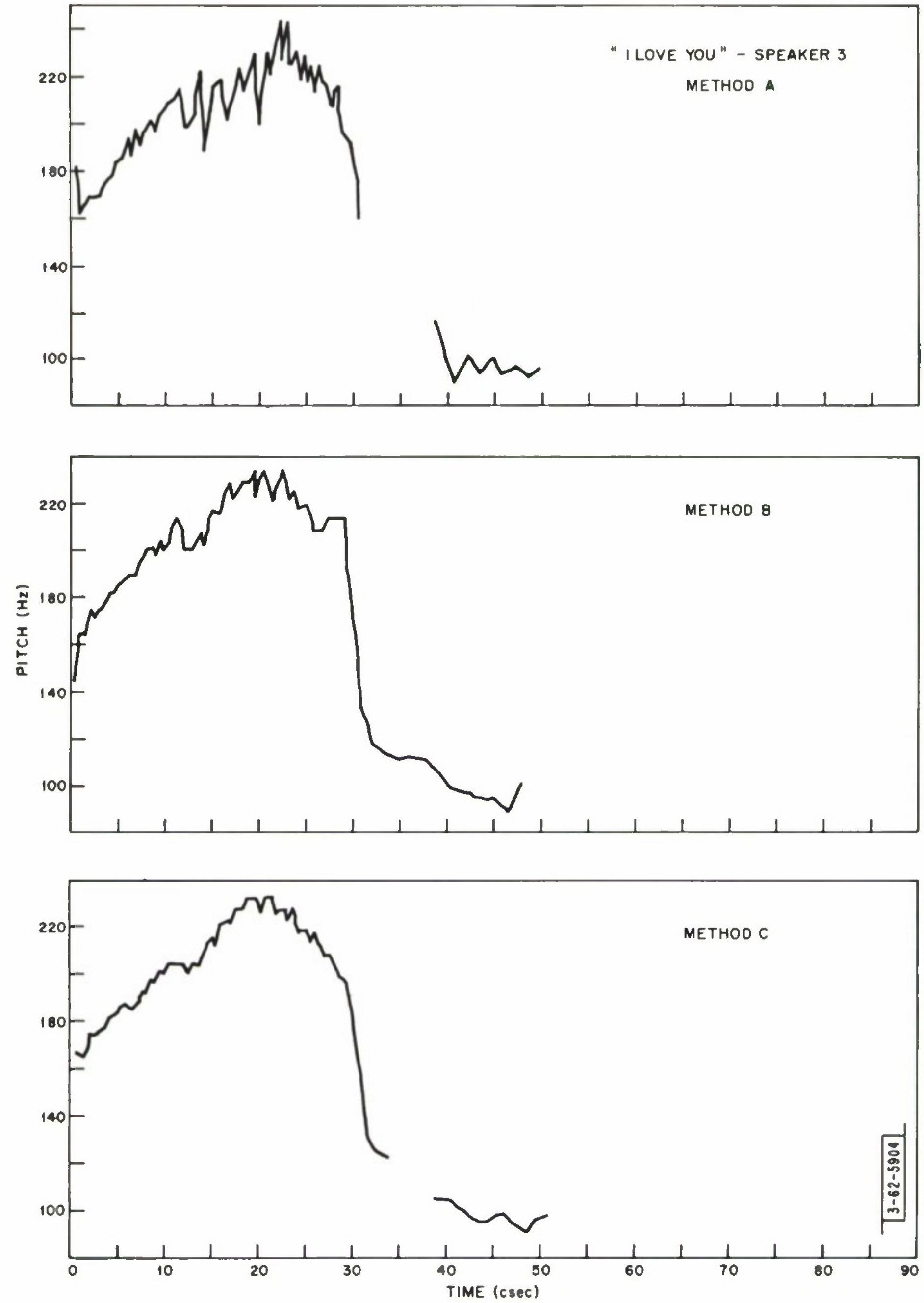

Fig. 17. Pitch Contours of Sentence "I Love You!" – Speaker 3.

APPENDIX 1
Description of Computer Facility

The computer simulations of this experiment were performed on MIT's Electrical Engineering PDP-1. This 18-bit machine is programmed for time sharing and has 2^{13} registers of accessible memory with another 2^{12} registers for the time sharing routines. Because of this time sharing facility it was relatively easy to get time on the PDP-1 for debugging the program. Since the actual simulation required some real-time operations the final simulation was performed in time sharing mode with all other users turned off. In this manner it was possible to use several of the time-sharing commands and yet have the full attention of the machine at all times.

Although it was necessary to use machine language mnemonics in programming the computer, many of the machine's other features made this machine nearly ideal. In addition to the core memory an additional 64K is supplied by a drum. Speech could be read into the computer using the Digital Equipment Corporation's 8-bit A-D converter and analog voltages could be obtained using its D-A converter (see Appendix 3). A cathode-ray oscilloscope provided easy visual observation of the data during many portions of the simulation and a light pen enabled further manipulation of this data.

The instruction list of the PDP-1 provided easy manipulations on the 18-bit memory words. The cycle time was $5\mu s$, but most instructions were at least two cycles long. The multiply instruction generally took about $30\mu s$. Since this instruction occurred often in the simulation, it accounted for the program taking approximately 100 times real time.

On-line communication with the machine could be performed in a variety of ways. This list included the typewriter console itself, a set of 18 toggle switches, 6 sense switches, a set of potentiometers controlling voltages to an A-D converter, three other A-D inputs, the light pen and finally punched paper tape. All these forms of inputs were used in the simulation.

At this time the PDP-1 is undergoing several modifications to add to this list of inputs and make it even more useful in speech work.

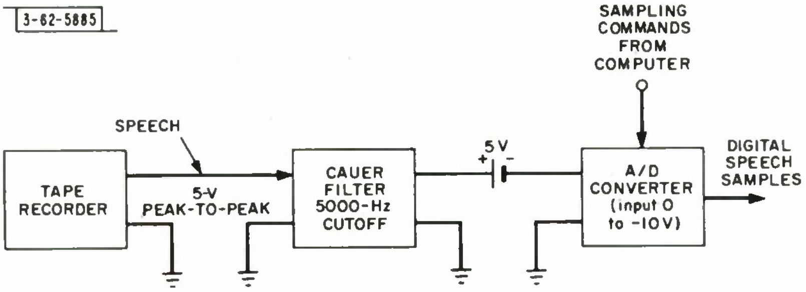
3-62-5885
SAMPLING COMMANDS FROM COMPUTER
SPEECH
TAPE RECORDER
5-V PEAK-TO-PEAK
CAUER FILTER 5000-Hz CUTOFF
5 V
+
-
A/D CONVERTER (input 0 to -10V)
DIGITAL SPEECH SAMPLES

Fig. 18. Analog-Digital Conversion of Speech.

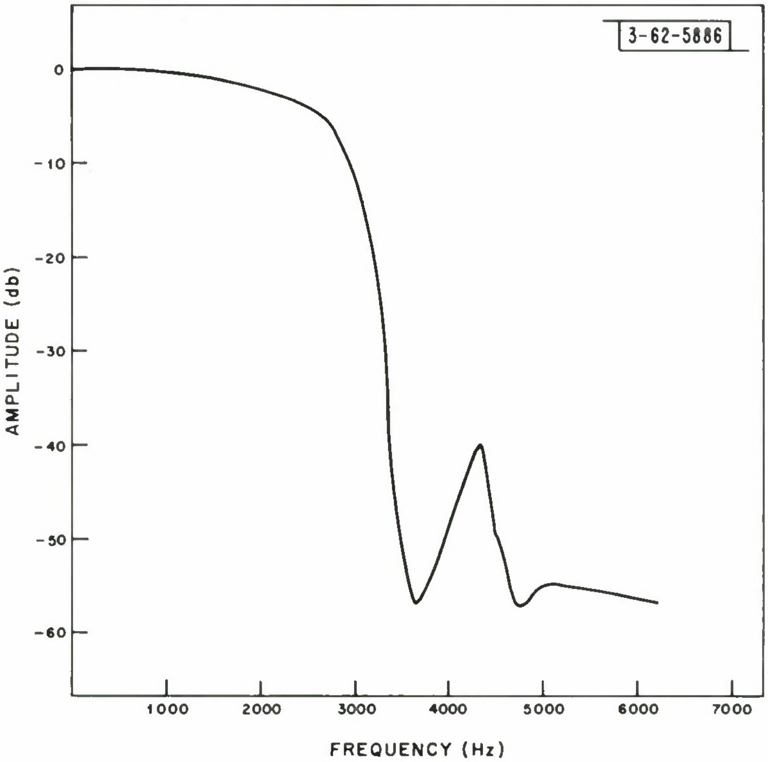

Fig. 19. Frequency Response of Analog Cauer Filter.

APPENDIX 2

Speech Input and Output on Computer

Before the simulation could be performed, the analog speech signal on the magnetic recording tape input had to be converted into digital data by an A-D converter. Only after this operation could the speech in the form of numerical data be stored in the memory of the PDP-1. Since the PDP-1 had an A-D converter associated with it, the entire digital conversion and data storage could be done in one program.

To understand this operation it is first necessary to know a little bit about the A-D converter. On command from the PDP-1 this converter will sample an analog signal between 0 and -10 volts and convert this sample into an unsigned 8-bit quantity. The computer will halt during this conversion, which takes $30\mu s$. In order to accomplish the A-D conversion, two things must then happen. First the input to the A-D converter must be between 0 and -10 volts and second the computer must be made to command the A-D to sample at the appropriate rate. This first requirement is easily met in Fig. 18. If we restrict the analog speech output of the tape recorder to be ± 5 volts, then, by biasing this with a -5 volt source in series with the tape recorder, the input to the A-D converter will be between 0 and -10 volts. To insure that the sampling rate of 10 kc/sec used in this simulation was at least twice as high as the highest frequency present in the input, a Cauer filter with frequency characteristics illustrated in Fig. 19 was cascaded with the tape recorder. This filter had to pass d.c. to insure the presence of the proper levels to the A-D converter.

The fulfillment of the second requirement is understood by reference to the flow chart of Fig. 20. The program is designed to sample 1-1/2 seconds of speech (15,000 samples at a 10 kc/sec rate) and store them by packing 2 to a word in the computer memory. The drum could not be used to store speech during read in because the drum write instruction took too long to execute. To insure that the desired 1-1/2 seconds of speech would be read in, a threshold detection scheme was used. If in the beginning the sample read in was not above a certain threshold, T, the sample would not be stored and a new sample would be taken. This would continue until a sample exceeded the threshold and then the remaining samples would be read in and stored in the machine. In this way the silence preceding the beginning of the utterance would not be stored in the computer. To achieve the current sampling rate, it was necessary to make use of the internal computer timing. This was done by designing the read-in program so that the total execution time of the instructions was precisely 100 μ seconds.

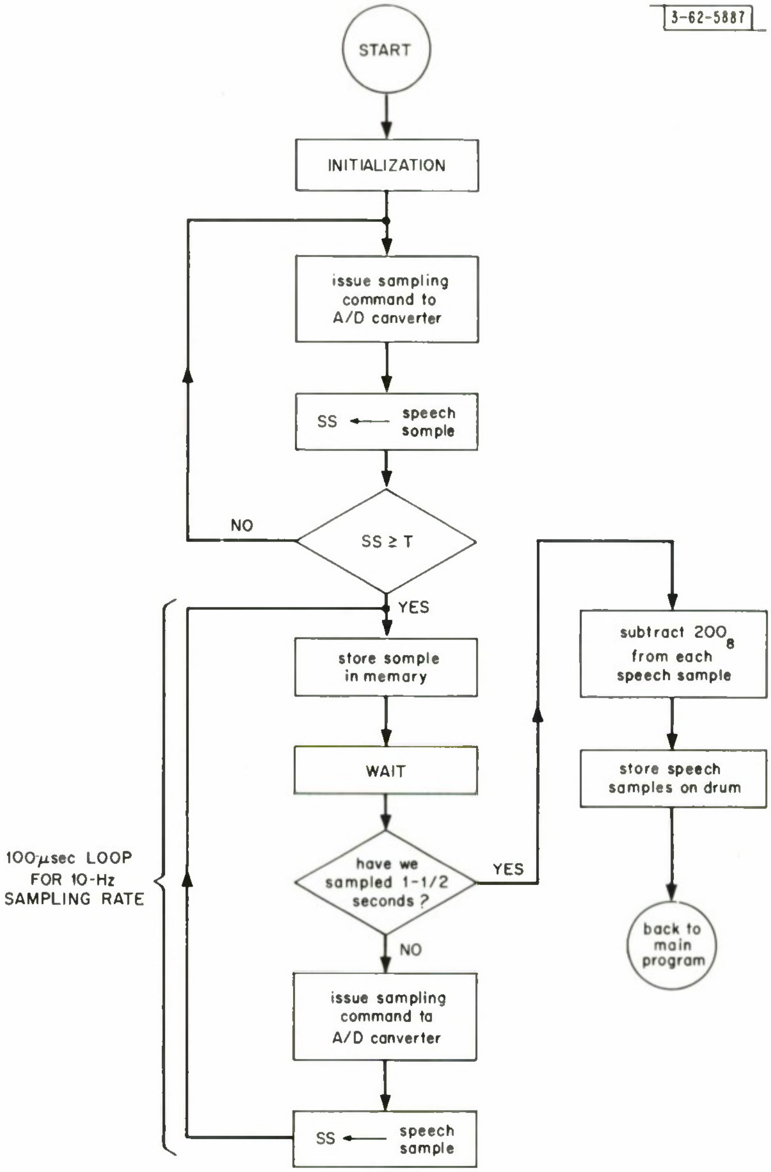
3-62-5887
START
INITIALIZATION
issue sampling command to A/D converter
SS ← speech sample
SS ≥ T
NO
YES
store sample in memory
WAIT
have we sampled 1-1/2 seconds ?
YES
NO
issue sampling command to A/D converter
SS ← speech sample
subtract 200₈ from each speech sample
store speech samples on drum
back to main program
100-μsec LOOP FOR 10-Hz SAMPLING RATE

Fig. 20. Flow Chart of Speech Read-In Program.

At the end of the sampling an additional program was necessary to convert the stored data into a form appropriate for input to the digital filters. Since the A-D converter presented the computer with a positive number between 0 and 377_8 it was necessary to subtract 200_8 from this data before use so that the data ranged betwen $+ 177_8$ and $- 177_8$ corresponding to the tape recorder output of $\pm$ 5 volts. This conversion was performed outside the sampling loop because of timing problems within the loop.

As a check on the speech data a second program using the D-A converter was written to convert this digital data into analog voltages at a 10,000 cycle/sec rate. This analog voltage was then passed through the Cauer filter discussed above, amplified and fed into a loudspeaker. Thus, it was possible to listen to the speech data stored in the machine. A further check could be made by visually displaying the speech data on the 'scope face.

Read Out

The final output from the computer simulation was a two-track tape recording. On one track was the analog speech, and on the other the pitch pulses representing the computed periods. This recording was made in steps illustrated in Fig. 21. The same speech which was originally read into the PDP-1 for determination of pitch was again fed into the A-D converter using the technique described earlier. This same signal was also recorded on the first track of the two-track tape. Again the analog speech was sampled by the PDP-1 and sent through the same threshold detection scheme used before to determine the beginning of the utterance (see flow chart of Fig. 22). This time, however, when the threshold was exceeded the computer, via the D-A converter, began to generate 5 volt, 50 μs wide pulses. These pulses which are made to occur once every computed pitch period of the wave are recorded on the second track. Note that the data recorded on the two tracks is being done by two separate means. They are brought into synchronization by having the computer note the same beginning of the utterance (assuming the threshold is the same as before and is high enough) as that data it has previously used to compute the pitch. The computer knows in advance which samples have a pitch indication associated with them and by asking every 100 μs (the same rate as the original read in) if a pitch pulse should be generated can now output the pulses at the proper time with respect to the recording of the first track.

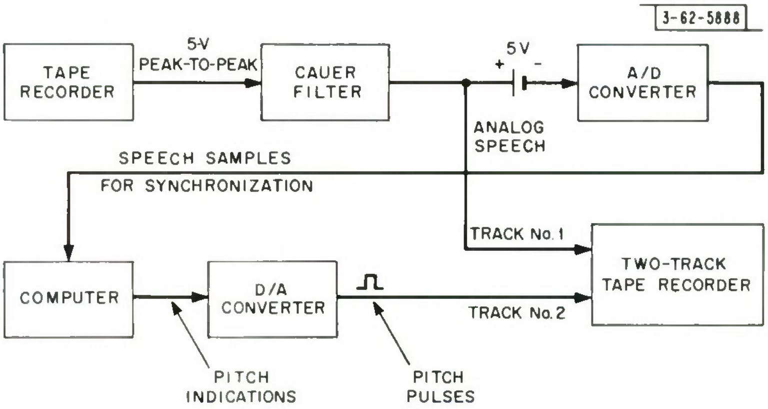

Fig. 21. System Used in Recording Speech and Pitch Pulses in Synchronization.

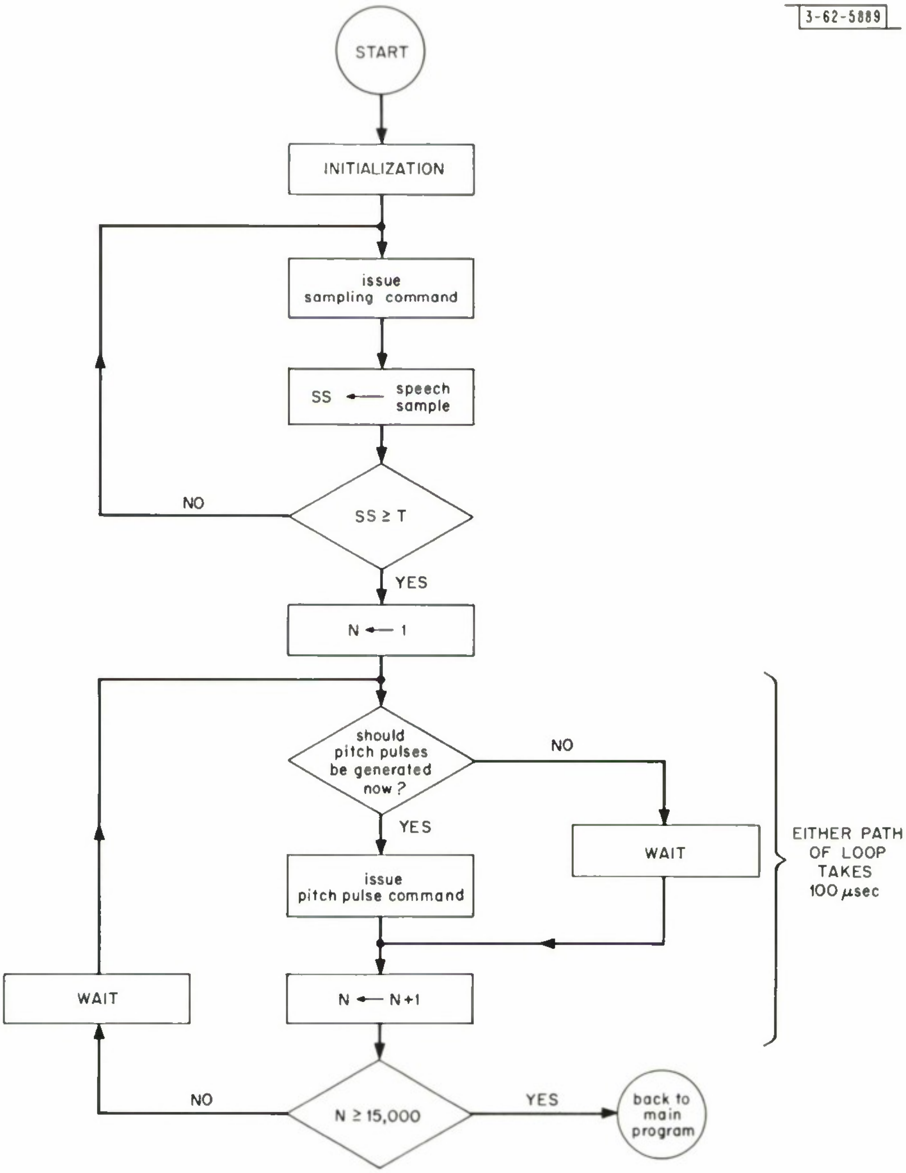

Fig. 22. Flow Chart of Pitch Read-Out Program.

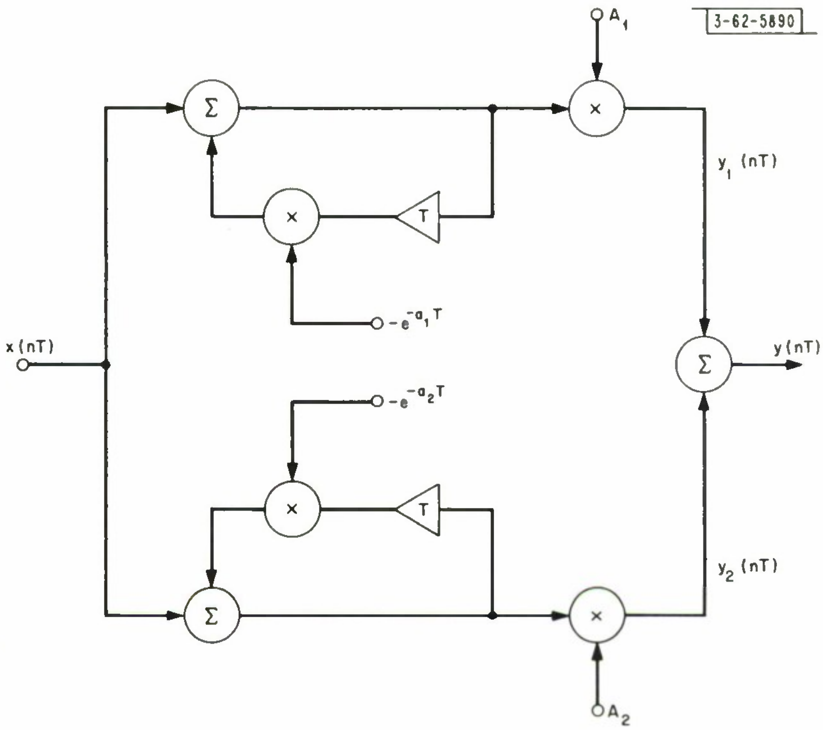

Fig. 23. Digital System Corresponding to a Continuous Time System Having an Impulse Response

$$h(t) = A_1\, e^{-a_1 t} + A_2\, e^{-a_2 t} \quad ; \quad t \geqslant 0$$

APPENDIX 3

Digital Filter Design Techniques

This appendix discusses the digital filters used in this experiment. All the filters used have direct analog counterparts. In fact, these counterparts gave rise to the digital filters by means of the methods of z-transforms.[14] By using the results of z transforms it was possible to go from the continuous time system having impulse response

$$h(t) = A e^{-at}$$

to the difference equation

$$y(nT) = A\, x(nT) - e^{-aT}\, y(nT - T)$$

where T is the sampling interval, which has an impulse response which is that of the samples of the impulse response of the continuous time system. For linear continuous filters wherein the impulse response can be expressed as a sum of exponentials, it is possible to get the desired digital system. For example, if

$$h(t) = A_1 e^{-a_1 t} + A_2 e^{-a_2 t}$$

then the digital system could be described by the equations

$$\begin{aligned} y_1(nT) &= x(nT) - e^{-a_1 T} y_1(nT - T) \\ y_2(nT) &= x(nT) - e^{-a_2 T} y_2(nT - T) \\ y(nT) &= A_1 y_1(nT) + A_2 y_2(nT) \end{aligned}$$

which is shown described pictorially in Fig. 23.

The 300-900 Hz filter used to eliminate the speech fundamental was the digital counterpart to a 7-pole Lerner bandpass filter[12]. Pictorially the filter is represented in Fig. 24. The constants representing pole positions of the digital filter were computed using the PDP-1. The frequency response and impulse response of this filter are depicted in Fig. 25 and Fig. 26 respectively.

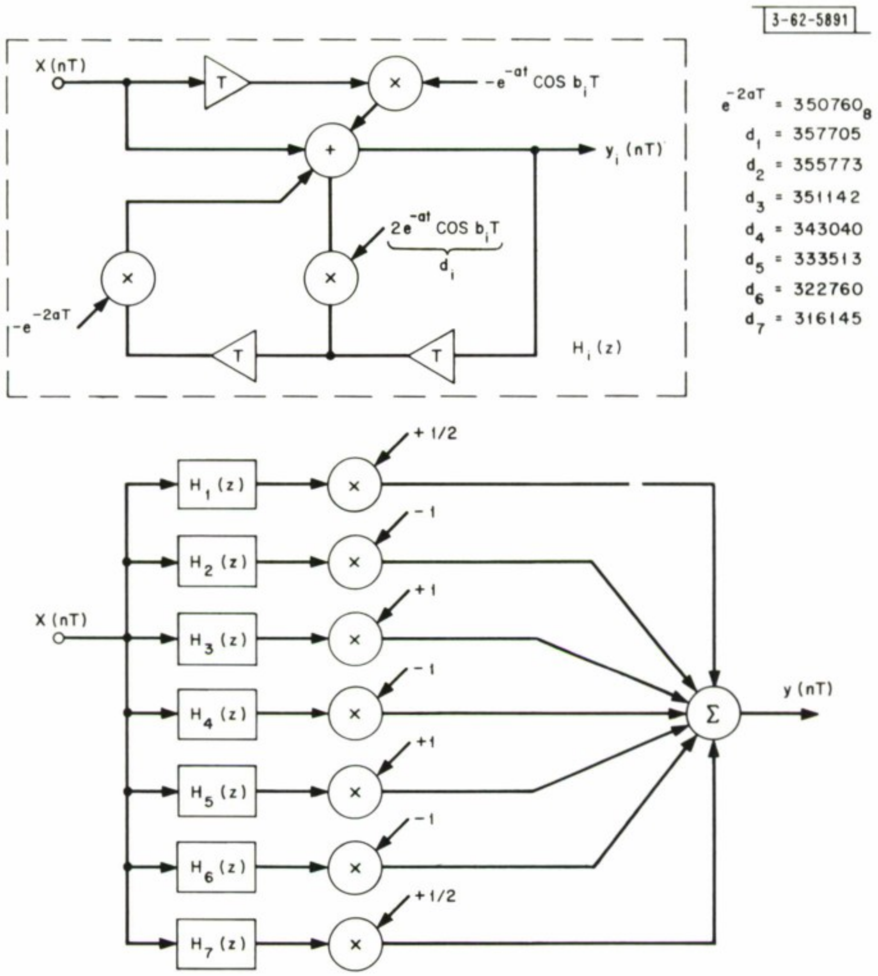

Fig. 24. Seven Pole Lerner Bandpass Filter in Digital Form.

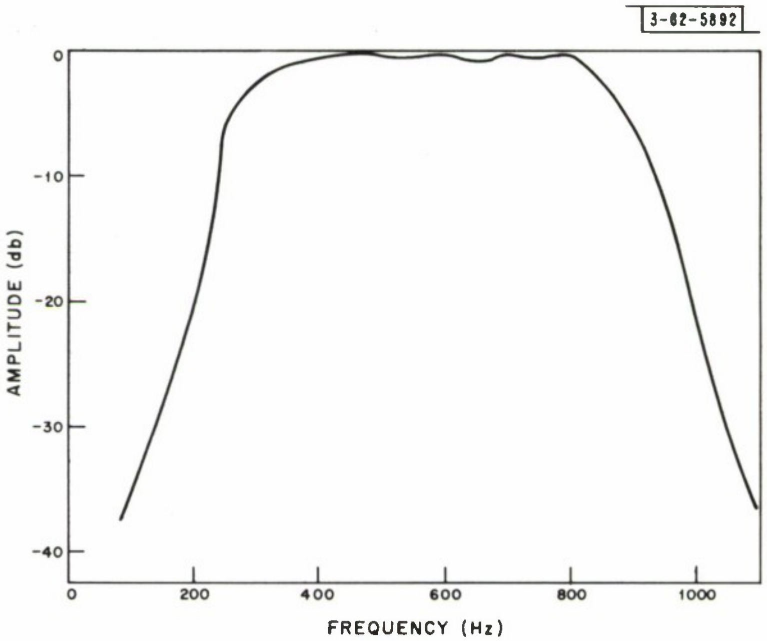

Fig. 25. Frequency Response of Seven Pole Lerner Filter.

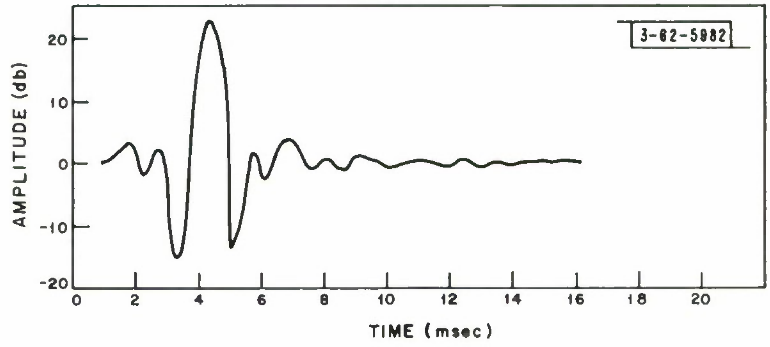

Fig. 26. Digital Impulse Response of Seven Pole Lerner Bandpass Filter (300-900 Hz).

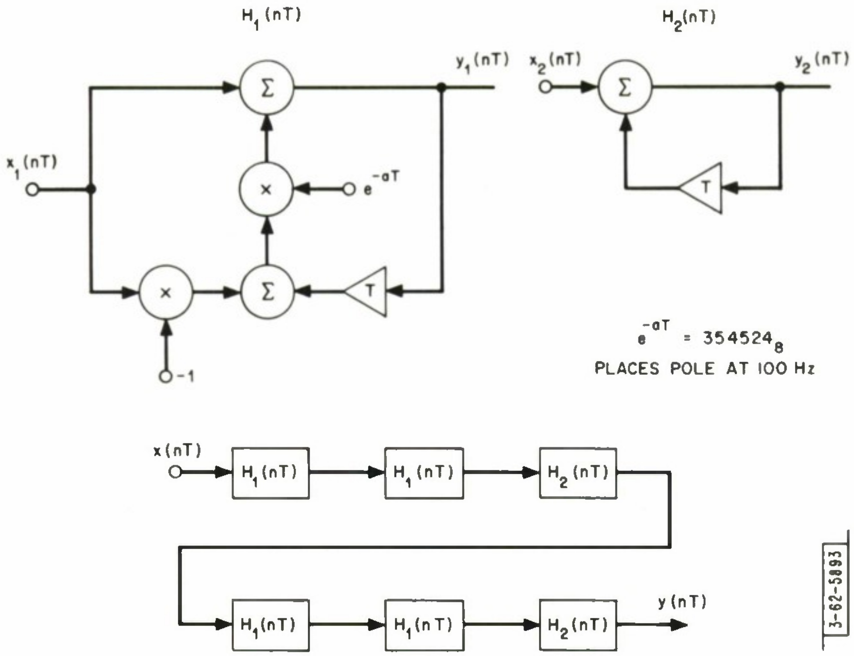

Fig. 27. Digital Representation of Slope Filter.

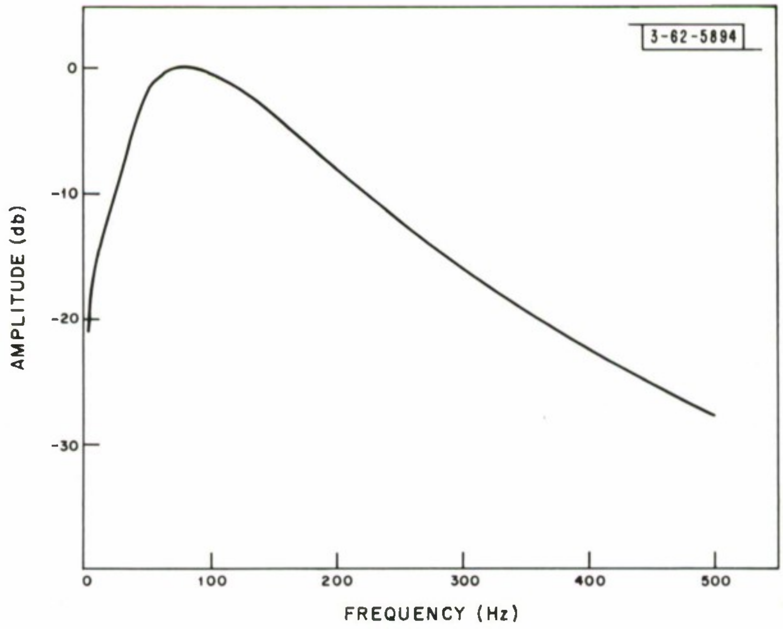

Fig. 28. Frequency Response of Slope Filter (12 db/Octave).

Similar digital methods were used to program the slope filter used in the experiment. Instead of arranging several digital systems in parallel as above, the systems were cascaded to produce the desired responses. See Fig. 27. Figure 28 and Fig. 29 are respectively the frequency and impulse responses of this filter.

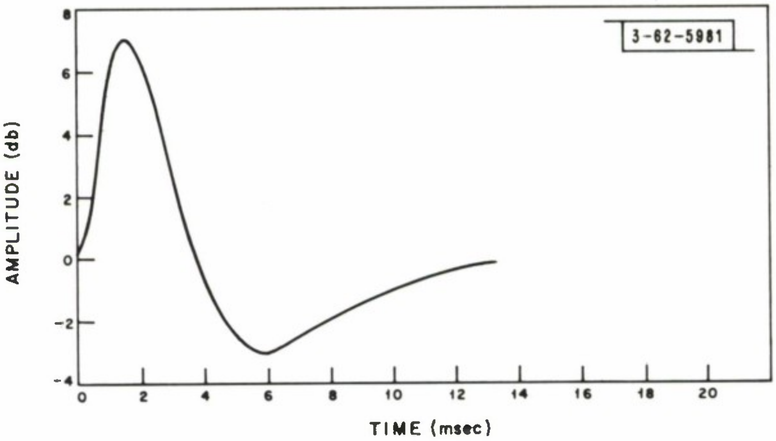

Fig. 29. Digital Impulse Response of Slope Filter (12 db/Octave Falloff).

ACKNOWLEDGMENT

The author wishes to acknowledge the encouragement and constructive criticism of Dr. Bernard Gold, his thesis supervisor.

J. Mackenzie and R. Butler of the Research Laboratory of Electronics provided valuable information about the computer hardware used in this thesis. The aid of J. Tierney and E. Aho of Lincoln Laboratory, in using their vocoder and auxiliary equipment is also greatly appreciated.

The work presented in this report was sponsored through a Graduate Traineeship from the National Science Foundation, Washington, D. C. and was performed under the direction of both the Research Laboratory of Electronics and Lincoln Laboratory of the Massachusetts Institute of Technology.

BIBLIOGRAPHY

1. H. Dudley, "Remaking Speech," J. Acous. Soc. of Am. 11, 169 (October 1939).

2. M. R. Schroeder, "Vocoders, Analysis and Synthesis of Speech," Proc. IEEE 54, 720 (May 1966).

3. J. L. Flanagan, *Speech Analysis and Perception* (Academic Press, Inc., 1965).

4. J. Tierney et al, "Channel Vocoder with Digital Pitch Extractor," J. Acous. Soc. of Am. 36, 1901 (October 1964).

5. H. Fujisaki, "Automatic Extraction of Fundamental Period of Speech by Autocorrelation Analysis and Peak Detection," J. Acous. Soc. of Am. 32, 1518 (November 1960).

6. J. S. Gill, "Estimation of Vocal Excitation During Speech with Particular Reference to the Requirements of Analysis-Synthesis Telephony - Ph. D Dissertation," University of London, London, England (1961).

7. B. Gold, "Description of Computer Program for Pitch Detection," Paper G-34 in Proc. 4th ICA, Copenhagen, 1962.

8. N. L. Daggett, "A Computer for Vocal Pitch Extraction," Technical Note 1966-3, Lincoln Laboratory, MIT (1966), ESD-TR-66-45.

9. A. M. Noll, "Short-Time Spectrum and Cepstrum Techniques for Vocal-Pitch Detection," J. Acous. Soc. of Am. 36, (February 1964).

10. H. L. Shaffer, "Information Rate Necessary to Transmit Pitch Period Durations of Connected Speech," J. Acous. Soc. of Am. 36, 1895 (October 1964).

11. N. P. McKinney, "Laryngeal Frequency Analysis for Linguistic Research - Thesis," Communication Sciences Laboratory, University of Michigan, Ann Arbor, Michigan (September 1965).

12. R. M. Lerner, "Band-Pass Filters with Linear Phase," Proc. IEEE 52, (March 1964).

13. R. R. Riesz, U. S. Patent No. 2,522,593 Bell Telephone Laboratories, Inc.

14. C. M. Rader and B. Gold, "Digital Filter Design Techniques," Proc. IEEE, to be published.

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)* Lincoln Laboratory, M.I.T.	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP None

3. REPORT TITLE

Vocoded Speech in the Absence of the Laryngeal Frequency

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Technical Note

5. AUTHOR(S) *(Last name, first name, initial)*

Goldberg, Aaron J. Gold, Bernard (Editor)

6. REPORT DATE 3 April 1967	7a. TOTAL NO. OF PAGES 42	7b. NO. OF REFS 14

8a. CONTRACT OR GRANT NO. AF 19 (628)-5167 b. PROJECT NO. 649L c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) Technical Note 1967-9 9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)* ESD-TR-67-237

10. AVAILABILITY/LIMITATION NOTICES

Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES None	12. SPONSORING MILITARY ACTIVITY Air Force Systems Command, USAF

13. ABSTRACT

Most pitch excited channel vocoders require the fundamental or laryngeal frequency of the input speech to be present if the output speech is to be of high quality. In order to determine if speech whose fundamental is absent can have its pitch accurately restored so as to be used as an input to a vocoder, a computer simulation was performed. The fundamental was restored by passing the speech through a fullwave rectifier followed by a slope filter. The accuracy of the pitch restoration of this method was compared with that of simply measuring the pitch of speech whose fundamental was present by slope filtering alone. A third pitch detection method, that of visually displaying the speech waveform and determining the pitch by eye, was also used as a comparison. Pitch contours of the three methods indicate that pitch restored by fullwave rectification and slope filtering has larger perturbations than pitch as detected by slope filtering alone. Both methods produced pitch contours having much larger perturbations than pitch determined visually.

Speech whose pitch was determined by the above three methods was used to excite the spectrally flattened Lincoln Laboratory Vocoder. Listening tests of the vocoder output indicate that pitch restored by fullwave rectification and slope filtering produced rougher sounding speech than speech whose pitch was detected by slope filtering alone, but both methods produced speech having considerably more audible roughness than that produced by visually detected pitch. Finally, the sophisticated pitch detector of the vocoder itself produced speech of quality comparable to that determined visually.

14. KEY WORDS

Vocoder
Speech compression
Pitch detection
Full-wave rectification
Slope filtering
PDP-1